Genomic Analysis of *Caenorhabditis elegans* Reveals Ancient Families of Retroviral-like Elements

Nathan J. Bowen and John F. McDonald

Genome Res. 1999 9: 924-935
Access the most recent version at doi:10.1101/gr.9.10.924

Genomic Analysis of *Caenorhabditis elegans* Reveals Ancient Families of Retroviral-like Elements

Nathan J. Bowen and John F. McDonald[1]

Department of Genetics, University of Georgia, Athens, Georgia 30602 USA

Retrotransposons are the most abundant and widespread class of eukaryotic transposable elements. The recent genome sequencing of *Caenorhabditis elegans* has provided an unprecedented opportunity to analyze the evolutionary relationships among the entire complement of retrotransposons within a multicellular eukaryotic organism. In this article we report the results of an analysis of retroviral-like long terminal repeat retrotransposons in *C. elegans* that indicate that this class of elements may be even more abundant and divergent than previously expected. The unexpected presence, in *C. elegans*, of an element displaying a number of characteristics previously thought to be unique to vertebrate retroviruses suggests an ancient lineage for this important class of infectious agents.

Retroelements (e.g., endogenous retroviruses and retrotransposons) are a major component of eukaryotic genomes. For example, >50% of the maize genome is made up of retroelements (SanMiguel et al. 1996), whereas >90% of the genome in certain plants such as wheat and pine have been reported to be comprised of retroelements (Flavell 1986). In animals, the figures are also impressive. For example, at least 15% of the *Drosophila* (Capy et al. 1998) and 35% of the mouse and human (Yoder et al. 1997) genomes have been estimated to be comprised of retroelements. The biological importance of retroelements extends from their being major causes of mutation (Green 1988; Miki 1998) and disease (Kazazian 1998) to being significant factors in genome evolution (McDonald 1993, 1998; Britten 1996; Miller et al. 1996).

In an effort to identify factors that have shaped the evolution of this important component of contemporary genomes, our laboratory (Jordan and McDonald 1998,1999a,b) and others (SanMiguel et al. 1996; Kim et al. 1998; Marin et al. 1998) have recently initiated a genomics approach toward the study of retroelement evolution. The ongoing genome sequencing of a variety of model experimental organisms and humans is providing an unprecedented opportunity to examine the patterns of molecular variation existing among the entire complement of retrotransposons residing in genomes. Analysis of this data can provide novel insights into the tempo and mode of retroelement evolution. In this article we present a phylogenetic analysis of long terminal repeat (LTR) retrotransposons present within the genome of *Caenorhabditis elegans*. Our results indicate that there are no less than 12 distinct families of LTR retrotransposons represented within the *C. elegans* genome. These include one novel family that displays many features characteristic of complex vertebrate retroviruses such as the spumaviruses and the lentiviruses to which human immunodeficiency virus, HIV, belongs. This unexpected finding suggests that infectious vertebrate retroviruses may have a remarkably long ancestry and may have been components of ancient eukaryotic genomes.

[1]Corresponding author.
E-MAIL mcgene@arches.uga.edu; FAX (706) 542-3910.

RESULTS AND DISCUSSION

Distinct Families of LTR Retrotransposons Are Detected in *C. elegans*

The CLUSTALX program of Thompson et al. (1997) was used to align the amino acid sequences of all *Cer* proteins (see Methods). *C. elegans* LTR retrotransposons can be grouped into 12 families or distinct lineages based on the amino acid sequence of their reverse transcriptase (RT). We have designated ≥90% amino acid sequence identity of RT to define the individual families and have named the new families *Cer2–Cer12* (*C. elegans* retrotransposon) in keeping with the nomenclature previously used by Britten (1997). Multiple elements within a family are designated by a dash followed by an additional number. For example, we have identified a new element that is 99.9% identical at the nucleotide level to the previously described *Cer1* and have designated it as *Cer1-1*. Interestingly, each *Cer* family consists of only one or two elements.

Structural Analysis

In all *C. elegans* LTR retrotransposons thus far identified, the Integrase (IN) is located 3′ to the RT domain of the element, which is characteristic of *Gypsy–Ty3* group LTR retrotransposons and retroviruses. Table 1 lists other distinguishing characteristics of *Cer* elements including LTR lengths and percent (%) nucleotide identities between the 5′ and 3′ LTRs. LTR identity can be used to estimate the time elapsed since the element transposed (Jordan and McDonald 1998, 1999b;

9:924–935 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

Table 1. Cer Element Characteristics

C. elegans retrotransposable element	Genomic clone	Chromosome	Inserted element length (bp)	Integrase location relative to RT	LTR length/ % identity	LTR end dinucleotides	Direct flanking repeat
Cer1	f4432/par3	III	8,865	3′	491/100	tg/ct	atca/atca
Cer1-1	y43d4	IV	8,865	3′	491/99.6	tg/ct	agtg/agtg
Cer2-0	r03d7	II	9,846	3′	645/99.2	tg/ct	attg/attg
Cer3-0	f58h7	IV	8,720	3′	424/100	tg/ct	ataa/ataa
Cer4-0	t23e7/f15g10	X	5,081	3′	164/99.4	tg/ta	atttc/atttc
Cer5-0	t03f1	I	5,082	3′	211/100	ag/ca	gttag/cttat
Cer5-1	k02a2/y91d1a	II	5,008	3′	61/96.7	at/tt	agtat/agtgc
Cer6	e03a3/t02c12	III	5,111	3′	218/100	tg/ca	ctaaa/ctaaa
Cer6-1	y102a5c	V	~7,230	3′	~300 <95%	not identified	not identified
Cer7	zc132	V	10,052	3′	327/99.7	tg/ta	cccag/cccag
Cer8	zk228/y51a2c/zk262	V	19,496	3′	300–600/5′ has large insertion or 3′ has deletion	tg/ca	gcccg/gcccg
Cer9	y43f4a	III	13,588	3′	517/99.8	tg/ca	acttc/acttc
Cer10	f07c7/y91b8a/c35b8	X	12,451	3′	648/99.7	tg/ca	aagtt/aagtt
Cer10-1	t23b12	V	not identified	3′	not identified	not identified	not identified
Cer11	t14g12	X	8,000 (coding)	3′	not identified	not identified	not identified
Cer12	f55c9/f21d9	V	12,067	3′	537/99.6	tg/ca	gaaac/gaaac

SanMiguel et al. 1998). With the exception of one element, *Cer5-1*, all elements with distinguishable LTRs displayed >99% nucleotide identity between 5′ and 3′ LTRs, indicating relatively recent transposition of the elements. Table 1 also lists the dinucleotides found at the end of the LTRs. Many LTR-retrotransposons and retroviruses universally begin and end with the dinucleotide inverted repeat (DIR) TG...CA (Dej et al. 1998). This DIR was found in nearly half of the *Cer* elements characterized. We also identified the direct flanking repeats that are the result of repair of the integration event. The flanking repeats of *Cer* elements consist of either 4 or 6 nucleotides with no apparent conservation between elements.

Cer Protein Domains

Figure 1, A–E, depicts the amino acid alignment of each protein domain from *Cer* elements and identifies conserved regions of each domain. The common cysteine array, Cys-X_2-Cys-X_4-His-X_4-Cys (CCHC) of the Nucleocapsid (NC) domain of Gag (group specific antigen) was found in all elements except *Cer12* (Fig. 1A). One or two copies of the CCHC motif have been detected previously in the Gag proteins of LTR-retrotransposons and retroviruses (Coffin et al. 1997). The CCHC motif binds zinc and is thought to be important for binding to the RNA genome during element or viral assembly. Although *Cer12* lacks the CCHC motif, it is unlikely that *Cer12* represents a nonfunctional family of LTR retrotransposons because the homology between its LTRs (99.6%) indicates recent transposition. Interestingly, the CCHC motif has also been reported to be absent in spumaviruses and the yeast LTR retrotransposons Ty1 and Ty2 (Coffin et al. 1997). Three imperfect CCHC motifs are present in the *Cer7–Cer10*, *Cer10-1*, and *Cer11* elements. The presence of three imperfect copies of the CCHC motif is also a feature of several non-LTR retrotransposons including Jockey, Doc, Het-A, and Tart-B1 of *Drosophila* (Capy et al. 1997).

The Protease domain (PR) is located downstream of the NC domain in LTR-retrotransposons and retroviruses. PR is synthesized as part of the Gag-Pro-Pol precursor, usually as the result of a frameshift or readthrough (termination) suppression after Gag protein synthesis. The amino acids between the NC and RT motifs of the *Cer* elements are homologous to previously identified retroviral proteases. In particular, the conserved "D,T/S,G" tripeptide of retroviral PR is present in many *Cer* elements (Fig. 1B) (Doolittle et al. 1989).

The Pol region of retrotransposons is made up of three enzymatic domains that are processed from the Pol precursor. Pol consists of the RT, RNase H, and IN domains. An alignment of the RT domain of *Cer* elements is shown in Figure 1C. The characteristic motifs of RT, as previously defined by Xiong and Eickbush (1988), are present in all *Cer* elements (Fig. 1C). *Cer* elements *7, 8, 9, 10, 10-1, 11,* and *12* (*Cer7–Cer12*) contain an unusual "YVDN" tetrapeptide motif at the active site of RT. The RNase H domains of the *Cer* elements were aligned, and previously identified motifs

(McClure 1991) were boxed and labeled (Fig. 1D). Finally, the previously identified HHCC motif and DDE domain (Capy et al. 1996) of IN are present in all *Cer* elements (Fig. 1E).

Position of *Cer* Elements in Existing RT Phylogeny *Within* Gypsy–Ty3 *Group Clade*

Because the RT domain has the slowest relative rate of change among all retroelement proteins (McClure et al. 1988), RT multiple sequence alignments (Xiong and Eickbush 1988; Doolittle et al. 1989) have been used to elucidate the evolutionary relationships among retroelements. Following characterization of the RT domain of the *Cer* elements, we made alignments with previously reported RT sequences (Fig. 2A). The most conserved regions of RT are shown boxed in Figure 2A.

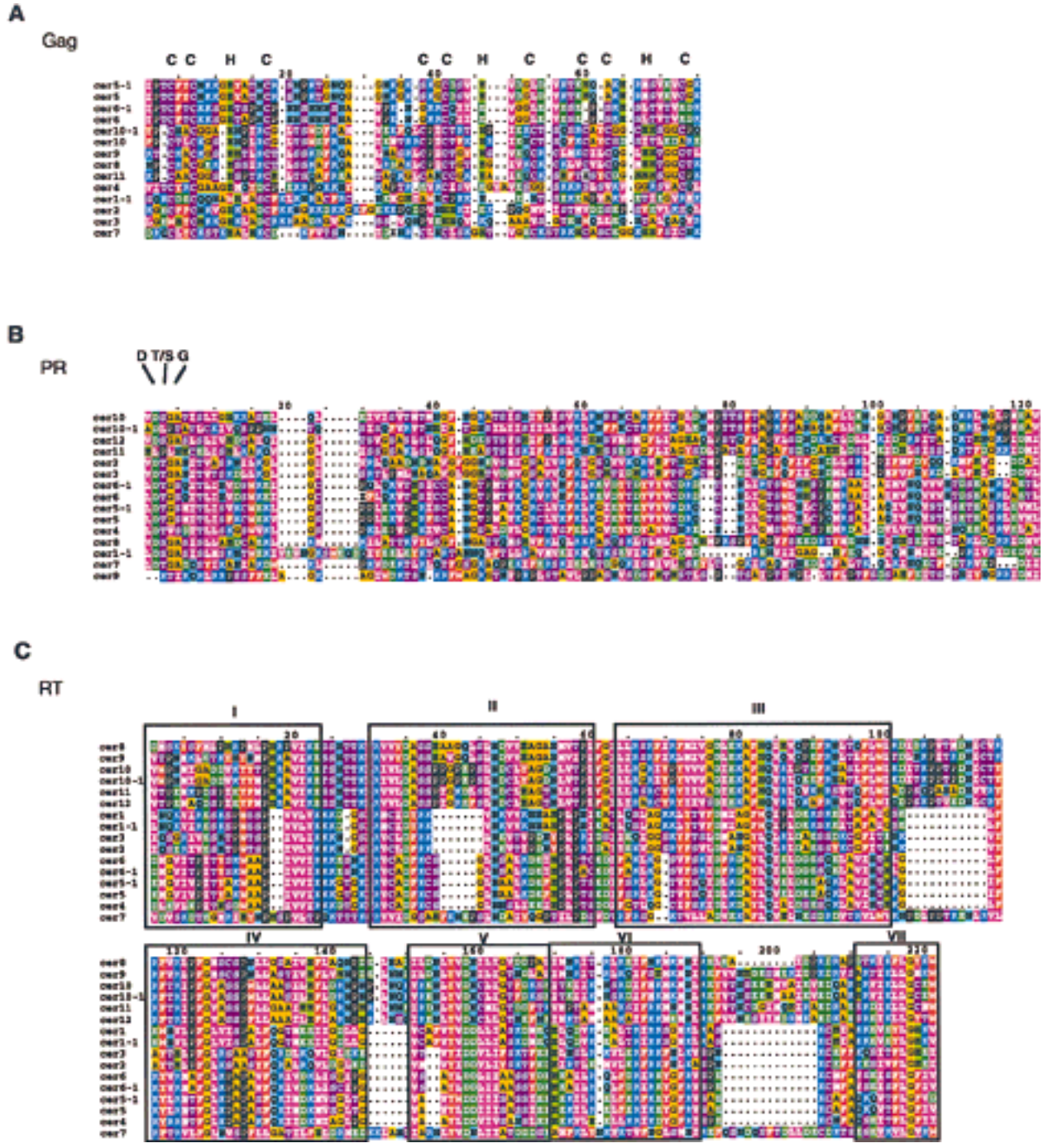

Figure 1 (*See facing page for D and* E *and legend.*)

These regions represent the RT ordered series of motifs (OSM) (Hudak and McClure 1999). The sequences comprising the RT OSM, as well as the full-length RT domain, were used to construct neighbor-joining (NJ) phylograms (Fig. 2B) from which the phylogenetic relationships of the *Cer* elements to other LTR-retrotransposons and retroviruses could be deduced.

The first group, to which *Cer1* belongs, clusters most closely to the *Gypsy–Ty3* group of LTR retrotransposons. Other elements clustering within the *Gypsy–Ty3* clade are *Cer2* and *Cer3*. These two elements group closest to the previously identified *SURL* element from echinoderms. *Cer4, Cer5, Cer5-1, Cer6*, and *Cer6-1* are most closely related to the previously identified retrotransposon, *Mag*, from *Bombyx mori* and like *Mag*, have relatively short LTRs (see Table 1).

Outside of Gypsy–Ty3 *Group Clade*

Cer7–Cer12 represents a unique branch from both the *Gypsy–Ty3* and the retroviral groups of retroelements. This branch of *Cer* elements is most closely related to previously described elements *Tas, Pao*, and *Bel* iso-

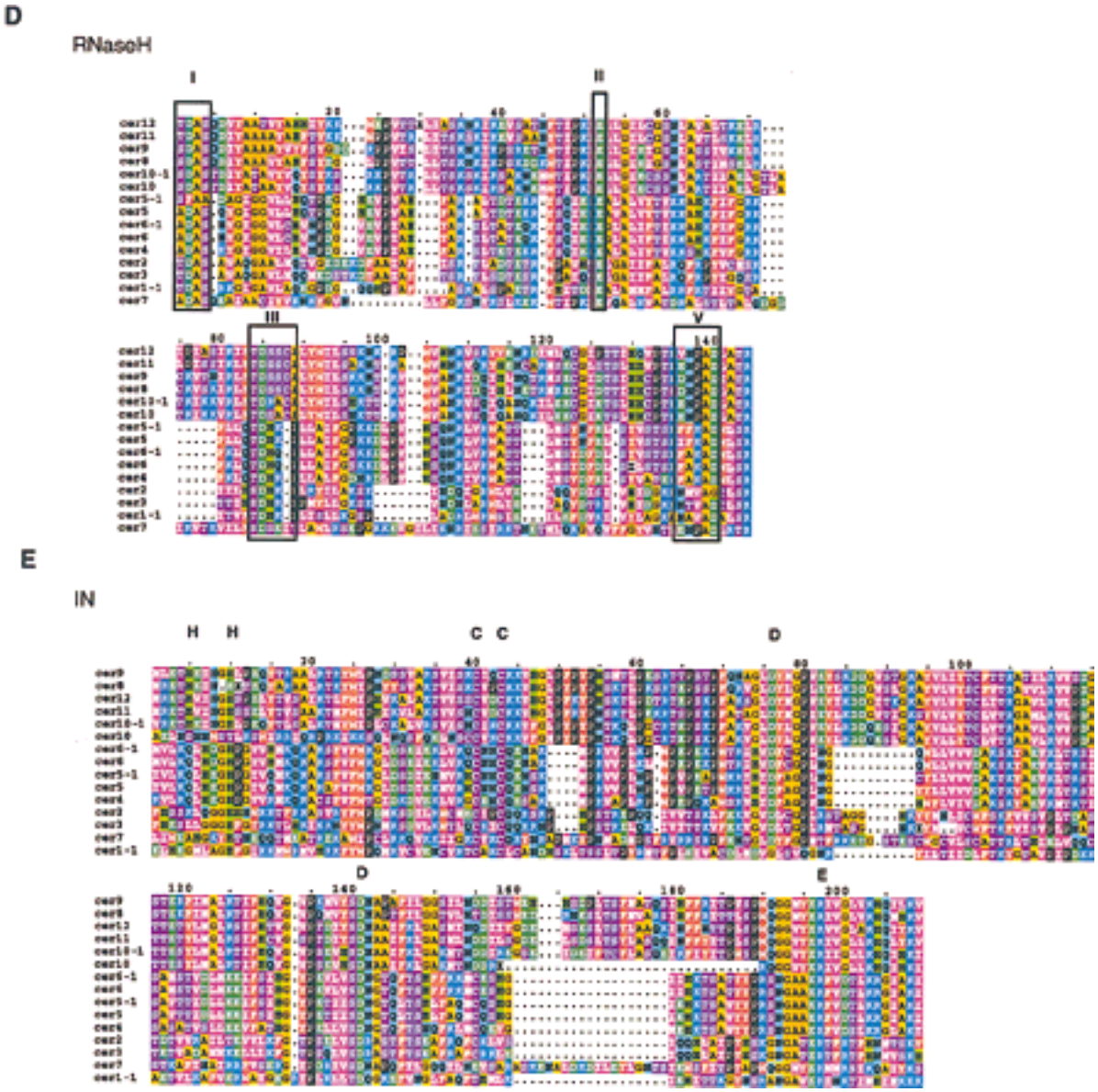

Figure 1 *Cer* protein alignments. The RT (*C*) of each *Cer* was identified according to the motifs previously designated by Xiong and Eickbush (1990). The additional protein domains (*A* = Gag, *B* = PR, *D* = Rnase H, and *E* = IN) were identified in a like manner using domains previously established by McClure (1991). The sequences were aligned using CLUSTAL X as described in Methods. Residue coloring was performed by MacBoxshade v2.01 and is based on the similarity scheme (F,W,Y), (I,L,M,V), (P), (D,E), (G,A), (S,T,C), (N,H),(R,K), (Q). Members of a similar residue group are the same color. In-frame termination codons or frameshifts are depicted by an X.

lated from *Ascaris lumbricoides*, *B. mori*, and *Drosophila melanogaster*, respectively (Xiong et al. 1993; Felder et al. 1994; Davis and Judd 1995). As noted above, one distinguishing characteristic of *Cer7–Cer12* is the presence of the conserved tetrapeptide YVDN at the active site of the RT enzyme. This motif was previously found in the RT region of the Tas element of *A. lumbricoides* (Felder et al. 1994). This motif differs from the tetrapeptide "F/YXDD" found in all previously reported RT sequences. Because the dipeptide "DD" had previously been thought to be essential for RT activity, it was speculated as to whether "YVDN" was indeed functional (Felder et al. 1994). Our finding that the "YVDN" motif is present in six distinct lineages of *Cer* elements argues strongly for its functionality. Additionally, all calculated LTR/LTR identities for *Cer7–Cer12* are >99%, indicating that they have recently transposed.

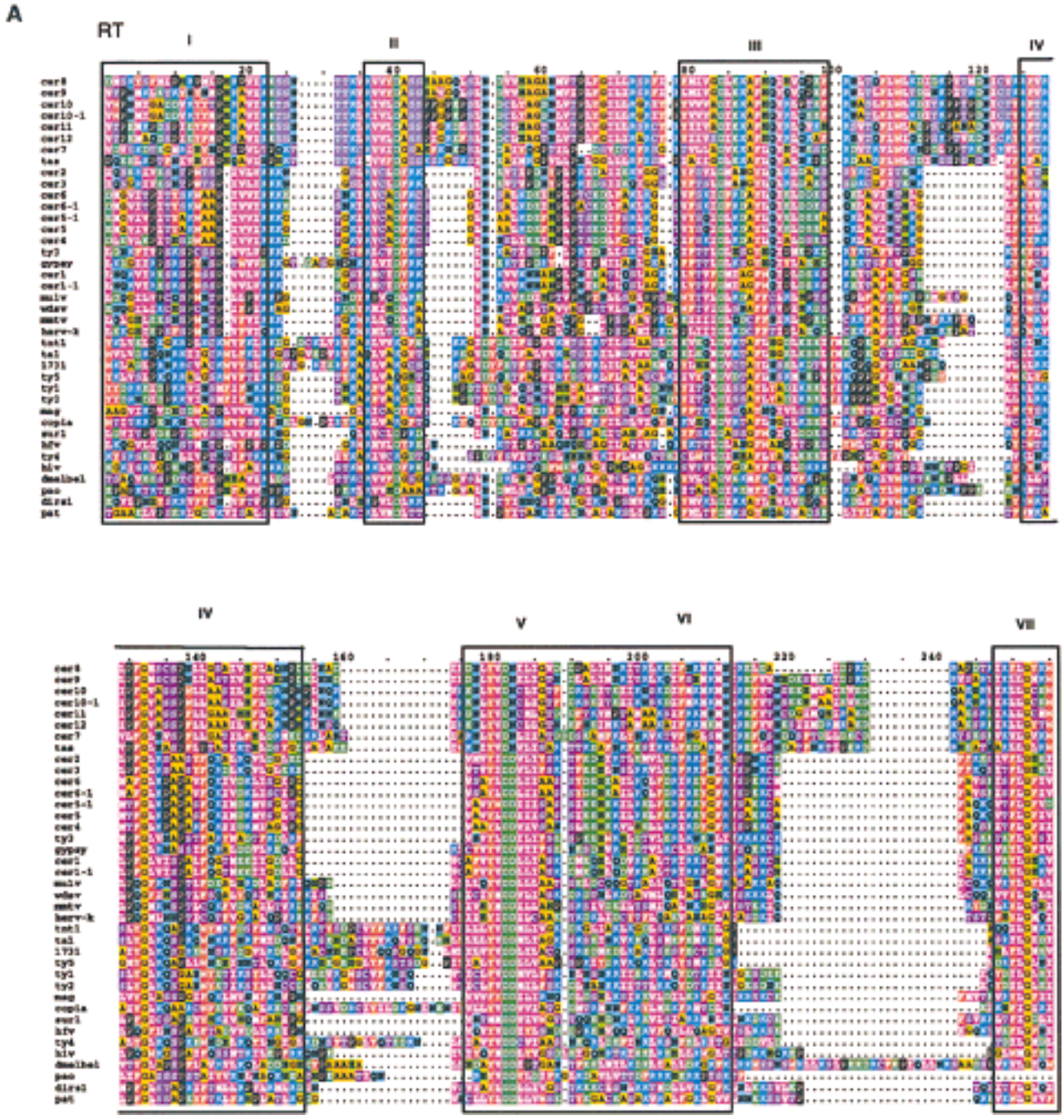

Figure 2 (*See p. 930 for legend.*)

One member of this group of elements, *Cer7*, contains a large ORF immediately following the IN domain that displays many features characteristic of the *envelope* (*env*) encoding region of retroviruses. The size and location of this ORF is nearly identical to that of the *env* encoding region of the *Acaris Tas* retroelement (Felder et al. 1994). The presence of putative *C. elegans* splice donor and splice acceptor sites in *Cer7* indicates that a smaller (subgenomic) RNA could be formed between the 5′ leader portion of the full-length (genomic) transcript and an acceptor site located upstream of the *env* initiation codon (Fig. 3). Although no

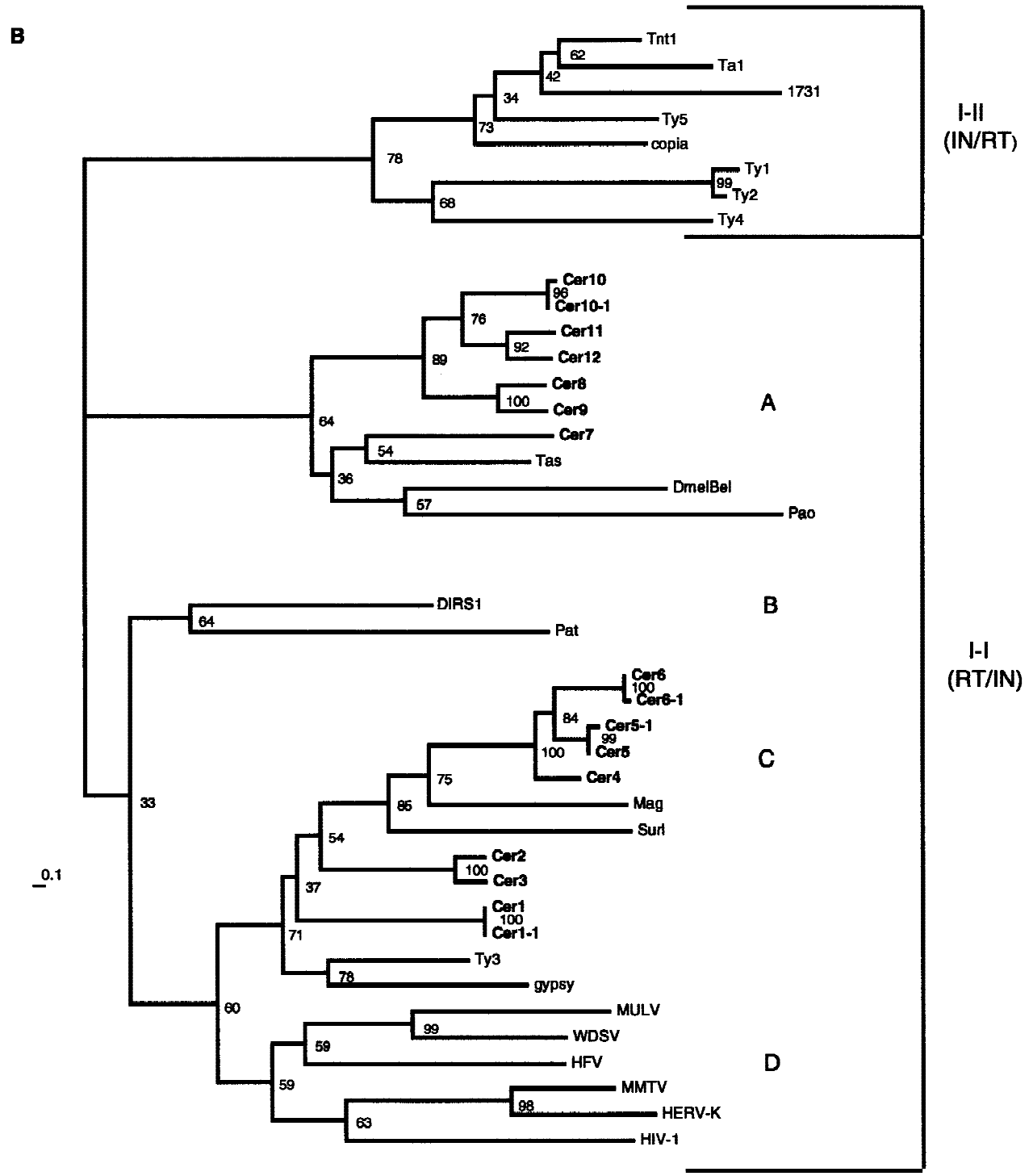

Figure 2 (*See p. 930 for legend.*)

significant sequence homology was found between the putative Env of *Cer7* and other retroviral Env proteins, this is to be expected given the nature of envelope function and the large divergence between host organisms. There are, however, many structural similarities between the putative Env of *Cer7* and other retroviral Env proteins. The putative *Cer7* Env contains a leader (L) signal peptide domain, multiple *N*-glycosylation sites, and a transmembrane region followed by a basic anchor (see Figs. 3 and 4A). Also present is a minimal furin-protease consensus cleavage site, "RXXR" (Molloy et al. 1992) (Figs. 3 and 4A), which may serve as the cleavage site between the integral surface glycoprotein (SU) and the transmembrane domain (TM). These are structural features common to all retrovirus envelope proteins (Coffin et al. 1997).

Another interesting feature of *Cer7* is that it has three small overlapping ORFs following its putative *env* gene (Fig. 3). This is a feature of the complex vertebrate retroviruses such as the human T-cell leukemia–bovine leukemia viral group (HTLV, BLV), the spumavirus, and the lentivirus groups. One of the small ORFs fol-

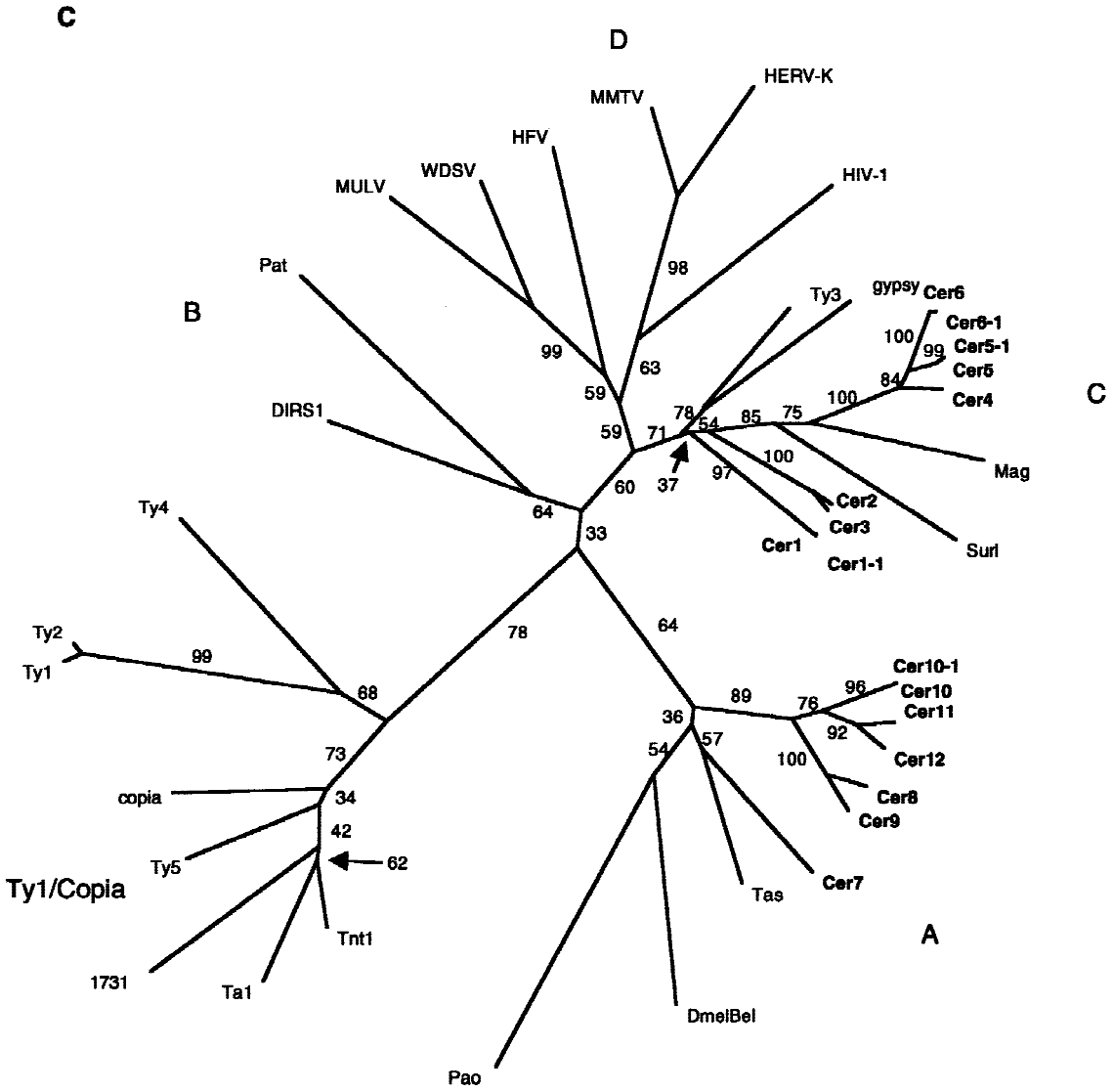

Figure 2 RT phylogenetic analysis. (*A*) RT amino acid alignment. The domains are boxed and numbered as by Xiong and Eickbush (1988). Additional RT sequences used in the Figure are referenced in Xiong and Eickbush (1990) or described as follows: (HFV) Human foamy virus, accession no. Y07725; (WDSV) walleye dermal sarcoma virus, accession no. AF033822; (PAT) *Panagrellus* redivivus retrotransposon, accession no. X60774; (Tas) *A. lumbricoides* retrotransposon, accession no. Z29712. Tas contains an in-frame termination codon at position 36 above. A gap was inserted at this position for the phylogenetic analysis. Residue coloring is based on the scheme described in Fig. 1. (*B*) NJ phylogram of sequences aligned in *A* (see methods). The tree was rooted with all of the *Copia–Ty1* group LTR retrotransposons. Major clades are labeled according to the nomenclature proposed in Table 2. (*C*) Unrooted NJ phylogram of sequences aligned in A. Major clades are labeled as in B.

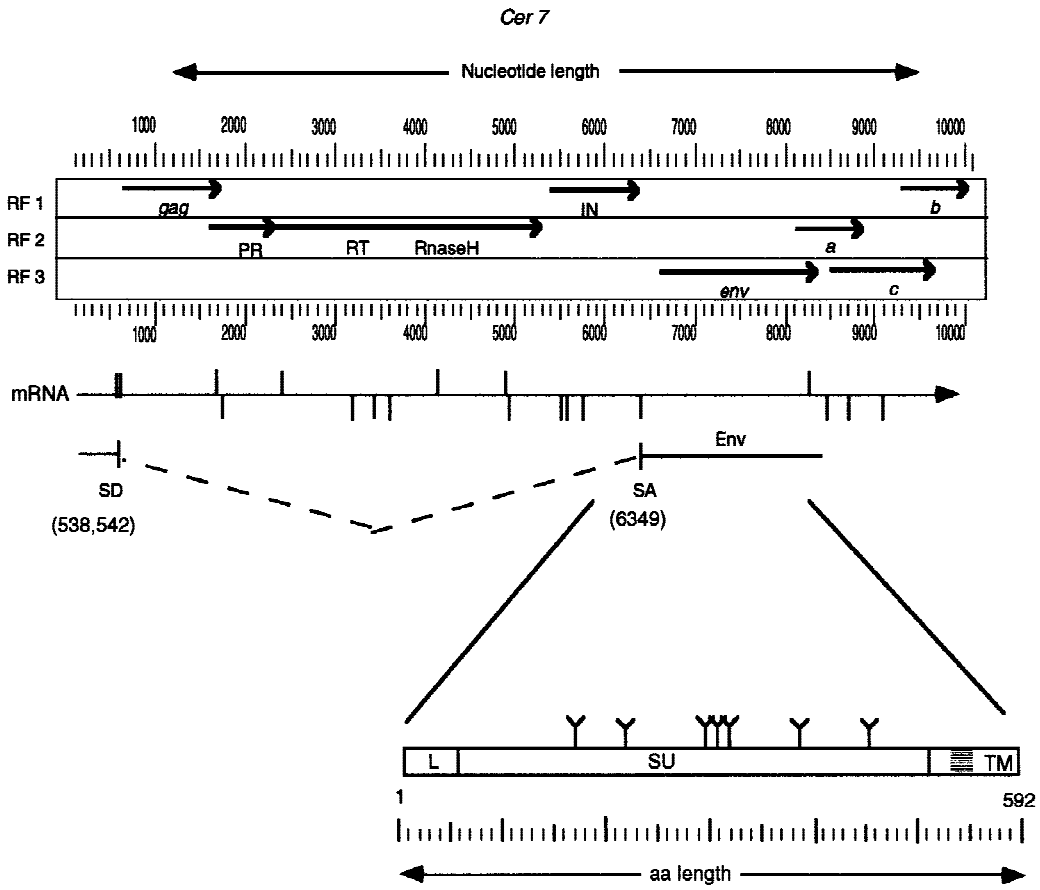

Figure 3 *Cer7* ORF map and putative Env. The largest ORFs (>200 amino acids) from the three forward reading frames (RF1, RF2, and RF3) of *Cer7* are translated and depicted as solid lines. Stop codons are shown as arrowheads. The common retroviral domains (PR and IN) and ORFS (*gag* and *env*) are labeled below the indicated ORFs. Note the three ORFs (a, b, and c) following the putative *env* ORF. Accessory proteins of complex retroviruses are found in analogous regions. The genomic mRNA is shown as a horizontal arrow below the ORF map. The predicted splice donor (SD) and splice acceptor (SA) sites are shown as vertical lines above and below the mRNA arrow, respectively. The predicted splice donor and acceptor sites that lead to the Env ORF production are shown below the mRNA. The Env protein is depicted as a box below the predicted spliced RNA. The leader peptide is indicated by an L. The possible *N*-glycosylation sites are depicted by a "Y" above the Env box. Vertical lines within the box represent putative protease cleavage sites that may serve to remove the leader peptide and cleave the region between the integral surface (SU) and transmembrane (TM) domains of Env. The predicted membrane anchor of TM is shown in grey. See Methods for a description of the functional motif predictions.

lowing the putative envelope of Cer7, which we call ORF C, has a strongly predicted *C. elegans* splice acceptor site ($D^2 = 13.77$) located upstream of its putative initiation codon. The predicted ORF C contains a region rich in cysteines and histidines, as well as two coiled–coil regions (Fig. 4B) that may function in DNA-binding or protein dimerization formation. These features are strikingly similar to other transactivating accessory proteins found in complex vertebrate retroviruses. For example, the HTLV-1Tax protein also contains a region rich in cysteines and histidines. Tax is thought to interact with bZip domains of ATF/CREB proteins to stably recruit CBP and other general transcription factors to the U3 region of the HTLV-1 LTR to remodel nucleosomes and activate transcription (Bex and Gaynor 1998). Likewise, spumaviruses encode proteins between their *env* and 3′ LTR (Bel-1 in human foamy virus and Taf in simian foamy virus) that stimulate transcription by interacting with multiple sites within the U3 regions of their LTRs. Although Tax, Bel1, and Taf have analogous functions, they are not homologous in sequence suggesting that these complex retrovirus functions may have arisen multiple times during evolution.

The presence of the retroviral-like *Cer7* within the *C. elegans* genome indicates that endogenous retroviruses may be quite ancient components of animal genomes. It has been previously hypothesized that retroviruses originated about the time of the mammals (Temin 1985; Doolittle et al. 1989). However, our findings suggest that the ancestor of vertebrate retroviruses may have had an early metazoan origin.

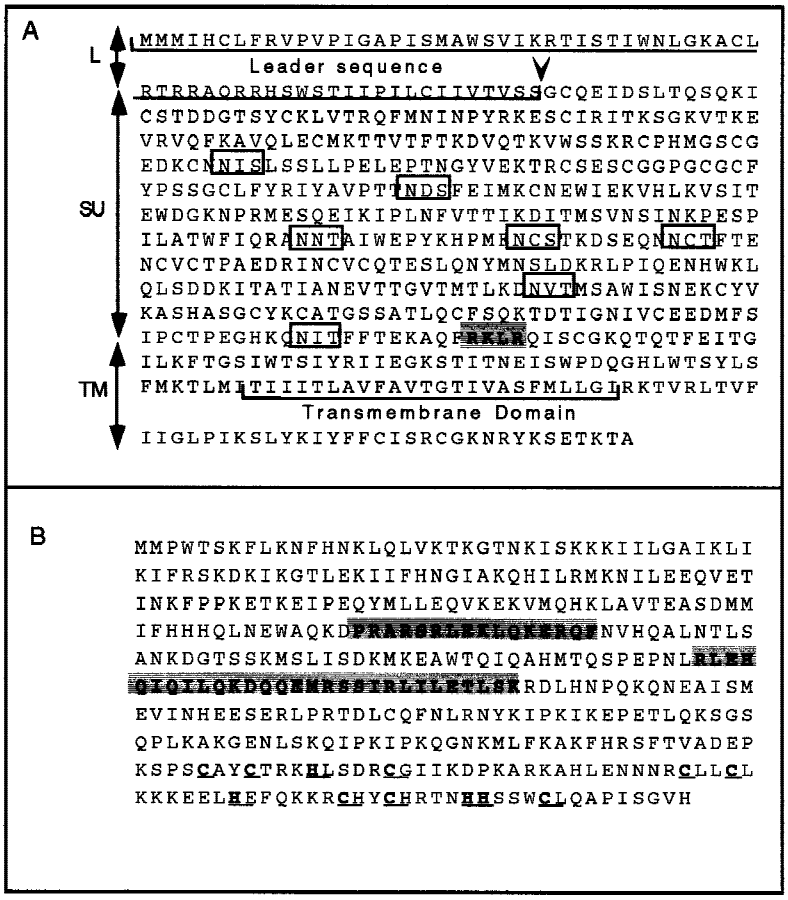

Figure 4 Amino acid sequences of putative Cer7 Env protein (*A*) and accessory protein C (*B*). (*A*) The leader sequence and transmembrane domain of Env are bracketed beneath and labeled. The predicted signal peptide cleavage site is indicated by a vertical arrowhead. The seven *N*-glycosylation targets "NXT/S" are boxed. The grey box indicates a putative furin cleavage site. Vertical double headed arrows at *left* of the sequence delineate the leader (L), surface (SU), and transmembrane (TM) domains of Env. (*B*) Gray boxes indicate the two predicted coiled-coil regions of ORF C. The mulitple cysteine and histidine residues are shown in bold and underlined. See Methods for a description of the functional motif predictions.

Retroelement phylogeny

In general, our results indicate that LTR-retrotransposons and retroviruses may be much more abundant and divergent than previously thought. The high level of sequence divergence among the *C. elegans* elements and previously characterized retroelement RTs indicates that a revision of LTR-retrotransposon and retrovirus phylogeny may be in order. An unrooted phylogenetic analysis of the sequences from Figure 2A places the clade containing *Cer7–Cer12* between the *Copia–Ty1* and *Gypsy–Ty3* groups (Fig. 2C). However, since the *Cer7–Cer12* clade is structurally similar to the *Gypsy–Ty3* group regarding the location of its *integrase* gene, it should be placed into the *Gypsy–Ty3* group. This creates a *Gypsy/Ty3* group with four major branches (Fig. 2B,C). The branches leading to these four major subgroupings as well as the branch leading to the *Ty1/Copia* group are supported by a majority of bootstrap replicates. These groupings/subgroupings include (1) the Ty1/Copia group, (2) the branch leading to the DIRS-1 and Pat elements, (3) the branch leading to the *Cer7–Cer12* elements as well as *Pao, Bel*, and *Tas*, (4) the branch leading to the group containing *Ty3* and *gypsy*, and finally, (5) the branch leading to the vertebrate retroviruses. Table 2 depicts a proposed nomenclature for the characterization of LTR-retrotransposons and retroviruses. Class I elements (i.e., those transposing via an RNA intermediate) can be divided first into three groups (I–III) based on defining structural differences. We propose group I to consist of LTR-retrotransposons and retroviruses with the RT/Rnase H/IN order of the *Gypsy–Ty3* group. Group II consists of LTR retrotransposons with the IN/RT/Rnase H order of the *Copia–Ty1* group, whereas group III consists of non-LTR retrotransposons. We propose that the next level of classification in group I be based on the major branches of the RT phylogeny. This level of classification is the least exact and is likely to undergo many additions and reclassifications as additional LTR retrotransposons are described that better delineate RT relationships. Based on this criterion, there are presently four major subgroups of group I. We designate these A–D as shown in Table 2. Subgroup A contains *Pao, Tas, Bel,* and *Cer7–Cer12*. Subgroup B contains the unusual retroelements *DIRS-1* from *Dictyostelium discoideum* that has inverted rather than direct terminal repeats and *PAT* from the nematode *Panagrellus redivivus* that has split direct repeats (de Chastonay et al. 1992). Subgroup C contains the well-known *Gypsy–Ty3* group, as well as *Mag* from *B. mori*. Subgroup D contains the vertebrate retroviruses. We propose that the next level of classification be based primarily on RT amino acid pairwise identity. An RT amino acid identity of ≥50% groups an element into the level of genus as used in the retroviral nomenclature (e.g., Lenti, Spuma, Type D, etc.). A ≥75% identity groups an element into a superfamily. Finally, a ≥90% identity groups an element into a family. Family members are designated by a dash followed by a number given upon order of identification. Names of genus, superfamily, and family are designated by the names of the first member of each group as indicated in Table 2.

The genome sequencing of model experimental

Table 2. Class I TE Nomenclature

Class (mode of transporation)	Group IN/RT order	Subgroup RT phylogeny	Genus >~50% RT amino acid ID	Superfamily >~75% RT amino acid ID	Family >~90% RT amino acid ID	Individual element
I (RNA)	I (RT/IN) *Gypsy–Ty3*	A	Pao			*Pao*
			BEL			*Bel*
			Tas			*Tas*
			Cer7			*Cer7.0*
			Cer8	Cer8	Cer8	*Cer8.0*
					Cer9	*Cer9.0*
			Cer10	Cer10	Cer10	*Cer10.0*
						Cer10.1
			Cer11	Cer11	Cer11	*Cer11*
				Cer12	Cer12	*Cer12*
		B	dirs-1	dirs-1	dirs-1	*dirs-1*
			PAT	PAT	PAT	*PAT*
		C	gypsy			*gypsy*
			Ty3			*Ty3*
			Cer1	Cer1	Cer1	*Cer1*
						Cer1.1
			Cer2	Cer2	Cer2	*Cer2*
					Cer3	*Cer3*
			Mag	Mag	Mag	*Mag*
			Cer4	Cer4	Cer4	*Cer4*
			Cer5	Cer5	Cer5	*Cer5*
						Cer5.1
			Cer6	Cer6	Cer6	*Cer6*
						Cer6.1
		D (vertebrate retrovirus)	Lenti			
			B-type			
			Spuma			
			MLV			
			Fish			
			D-type			
			ASLV			
			BLV–HTLV			
	II (IN/RT) *Copia–Ty1*		Ty1			*Ty1*
			Copia			*Copia*
	III (Non-LTR)					

organisms and humans is providing an unprecedented opportunity to examine the evolutionary relationships that exist among retroelements including LTR-retrotransposons and retroviruses. Analysis of LTR retrotransposons in the *C. elegans* genome indicates that this group of retroelements may be more abundant and divergent than previously suspected. In addition, the presence in *C. elegans* of elements displaying a number of characteristics previously thought to be unique to vertebrate retroviruses suggests an ancient lineage for this important class of infectious agents.

METHODS

Sequence Identification and Retrieval

The *C. elegans* genomic sequence data used in our analysis was accessed directly from the web sites of the Sanger Center, Hinxton Hall, Cambridge, UK, and the Genome Sequencing Center at the Washington University School of Medicine, St. Louis, Missouri Sequence retrieval was initiated by performing TBLASTN searches against the *C. elegans* genomic sequence (http://www.sanger.ac.uk/Projects/C elegans/blastserver.shtml) using the RT from the sole LTR retrotransposon representative from the *C. elegans* genome, *Cer1* (accession no. U15406) (Britten 1997). We also queried the current WORMPEP (http://www.sanger.ac.uk/Projects/C elegans/wormpep) database for ORFs with predicted homology to RT. Additional BLAST and TBLASTN searches were performed with the putative RTs until overlapping hits were retrieved by all sequences. A number of RTs belonging to the non-LTR containing retrotransposons were also detected within the *C. elegans* genome. These have been described previously by Marin et al. (1998).

Complete clones were retrieved from GenBank and characterized using SeqLab: The Graphical User Interface to the Wisconsin Package, (GCG 1999) maintained and made accessible by the Research Computing Resource (RCR) at the University of Georgia (UGA) (http://www.rcr.uga.edu/ biosci/home.html). Many elements were found to span two adjoining cosmids and were first aligned and made contiguous using GAP and BESTFIT. The dot matrix program COMPARE was used to identify regions of identity within each *C. elegans* cosmid. DOTPLOT was used to visualize the dot matrixes generated with COMPARE. LTRs appeared as two lines parallel to the identity diagonal (one above and one below). Subsequent analysis revealed the phylogenetically conserved TG...CA di-

nucleotide at the ends of many LTRs. Clone sequence, position, and ORFs were also viewed in A *Caenorhabditis elegans* Database (ACEDB) (Durbin and Thierry Mieg 1991) (documentation, code, and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk, and ncbi.nlm.nih.gov).

Multiple Sequence Alignments

The CLUSTAL X program (Thompson et al. 1997) was used to align the amino acid sequences of all *Cer* proteins and domains presented in Figure 1. RT amino acid sequences presented in Figure 2A were also aligned using the CLUSTAL X program. CLUSTAL X was run using the Gonnet 250 pairwise aligment and Gonnet series multiple alignment parameter settings. The Gonnet 250 and Gonnet series settings were able to recognize and align parts of all of the conserved regions of RT. These regions have been described previously (Xiong and Eickbush 1990; McClure 1993) and are collectively known as the RT OSM as described by Hudak and McClure (1999). The regions between the motifs are less conserved and represent hypervariable regions of the RT domain. The larger gaps introduced by CLUSTAL X were in these hypervariable regions. Using SeqLab, manual adjustments were made around regions containing gaps to minimize mutation events and to agree with previously published multiple alignments of RT (Xiong and Eickbush 1990; McClure 1991, 1992, 1993). Three different multiple sequence alignments were used in subsequent phylogenetic analyses. One analysis consisted of the entire RT, whereas a second consisted of only residues of the RT OSM (boxed in Fig. 2A). This served to eliminate the hypervariable and ambiguous regions of the alignments. Finally, a third alignment consisted of the regions between the RT OSM. These regions are also known as motif-intervening regions (MIRs).

Phylogenetic Analysis

Phylogenetic analyses were performed on the multiple sequence alignments using distance methods used by CLUSTAL X and PHYLIP (Felsenstein 1993). Draw N-J Tree and Bootstrap N-J commands of CLUSTAL X were used to generate nonbootstrapped and bootstrapped trees, respectively. The PRODIST program of PHYLIP using the Categories model was used to generate distance matrices that were analyzed with the NEIGHBOR program to generate NJ tree files. SEQBOOT was also used to generate 100 data replicates that which were subsequently analyzed with PRODIST (Categories model), followed by NEIGHBOR, and finally with CONSENSE to generate an unrooted bootstrapped tree. Analyses using the RT OSM, the MIRs, or the entire RT domain converged on an unrooted tree with the same general topology. This indicates a strong signal-to-noise ratio with the full-length multiple sequence alignment. We report the trees and bootstrap values generated with the PHYLIP package using the entire RT domain. The additional information included in the MIRs serves to strengthen the accuracy of the within subgroup phylogenetic reconstructions. The phylogram presented in Figure 2B was rooted with the Ty1/Copia group. All trees generated were visualized with TreeViewPPC version 1.5.3, (Page 1996).

Cer7 Protein Predictions

C. elegans splice sites were predicted by the SPL program of Baylor College of Medicine's GeneFinder (Solovyev et al. 1994). The leader peptide of the *Cer7* Env was predicted by SignalP version 1.1 (Nielsen et al. 1997). *N*-glycosylation sites were located by PROSITE (Hofmann et al. 1999). The transmembrane domain was identified by PHDhtm (Rost et al. 1995). Finally, the coiled-coil regions of *Cer7* C were predicted by COILS (Lupas et al. 1991).

RT Pairwise Identities

Pairwise amino acid identity spanning the entire length of the RT sequence shown in Figure 2A was used for the *Cer* family characterizations. PAUP 4.0b2a (Swofford 1999) was used to calculate pairwise mean differences that were converted to percent identities.

ACKNOWLEDGMENTS

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Bex, F. and R.B. Gaynor. 1998. Regulation of gene expression by HTLV-I Tax protein. *Methods* **16:** 83–94.

Britten, R.J. 1996. DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl. Acad. Sci.* **93:** 9374–9377.

———. 1997. Mobile elements inserted in the distant past have taken on important functions. *Gene* **205:** 177–182.

Capy, P., R. Vitalis, T. Langin, D. Higuet, and C. Bazin. 1996. Relationships between transposable elements based upon the integrase-transposase domains: Is there a common ancestor? *J. Mol. Evol.* **42:** 359–368.

Capy, P., T. Langin, D. Higuet, P. Maurer, and C. Bazin. 1997. Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? *Genetica* **100:** 63–72.

Capy, P., C. Bazin, T. Langin, and D. Higuet. 1998. *Dynamics and evolution of transposable elements*. Springer, New York, NY.

Coffin, J.M., S.H. Hughes, and H.E. Varmus. 1997. *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Davis, P.S. and, B.H. Judd. 1995. Nucleotide sequence of the transposable element, BEL, of Drosophila melanogaster. *Drosoph. Inf. Serv.* **76:** 134–136.

de Chastonay, Y., H. Felder, C. Link, P. Aeby, H. Tobler, and F. Muller. 1992. Unusual features of the retroid element PAT from the nematode Panagrellus redivivus. *Nucleic Acids Res.* **20:** 1623–1628.

Dej, K.J., T. Gerasimova, V.G. Corces, and J.D. Boeke. 1998. A hotspot for the Drosophila gypsy retroelement in the ovo locus. *Nucleic Acids Res.* **26:** 4019–4025.

Doolittle, R.F., D.F. Feng, M.S. Johnson, and M.A. McClure. 1989. Origins and evolutionary relationships of retroviruses. *Q. Rev. Biol.* **64:** 1–30.

Durbin, R. and J. Thierry Mieg. 1991. A *C. elegans* database. Documentation, code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov.

Felder, H., A. Herzceg, Y. de Chastonay, P. Aeby, H. Tobler, and F. Muller. 1994. Tas, a retrotransposon from the parasitic nematode Ascaris lumbricoides. *Gene* **149:** 219–225.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) v. 3.5c. Department of Genetics, University of Washington, Seattle, WA.

Flavell, R.B. 1986. Repetitive DNA and chromosome evolution in plants. *Phil. Trans. R. Soc. Lond. B. Biol. Sci.* **312:** 227–242.

GCG. 1999. (Genetics Computer Group). Wisconsin Package v.10.0. Genetics Computer Group, Madison, WI.

Green, M.M. 1988. Mobile DNA elements and spontaneous gene mutation. In *Eukaryotic transposable elements as mutagenic agents* (ed. M.E. Lambert, J.F. McDonald, and I.B. Weinstein), pp. 41–50. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Hofmann, K., P. Bucher, L. Falquet, and A. Bairoch. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27:** 215–219.

Hudak, J. and M.A. McClure. 1999. A comparative analysis of computational motif-detection methods [In Process Citation]. *Pac. Symp. Biocomput.*4: 138–149.

Jordan, I.K. and J.F. McDonald. 1998. Evidence for the role of recombination in the regulatory evolution of Saccharomyces cerevisiae Ty elements. *J. Mol. Evol.* **47:** 14–20.

———. 1999a. Phylogenetic perspective reveals abundant Ty1/Ty2 hybrid elements in the Saccharomyces cerevisiae genome *Mol. Biol. Evol.* **16:** 419–422.

———. 1999b. Tempo and mode of Ty element evolution in saccharomyces cerevisiae. *Genetics* **151:** 1341–1351.

Kazazian, H.H., Jr. 1998. Mobile elements and disease. *Curr. Opin. Genet. Dev.* **8:** 343–350.

Kim, J.M., S. Vanguri, J.D. Boeke, A. Gabriel, and D.F. Voytas. 1998. Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* **8:** 464–478.

Lupas, A., M. Van Dyke, and J. Stock. 1991. Predicting coiled coils from protein sequences. *Science* **252:** 1162–1164.

Marin, I., P. Plata-Rengifo, M. Labrador, and A. Fontdevila. 1998. Evolutionary relationships among the members of an ancient class of non-LTR retrotransposons found in the nematode Caenorhabditis elegans. *Mol. Biol. Evol.* **15:** 1390–1402.

McClure, M.A. 1991. Evolution of retroposons by acquisition or deletion of retrovirus-like genes. *Mol. Biol. Evol.* **8:** 835–856.

———. 1992. Sequence analysis of eukaryotic retroid proteins. *Mathl. Comput. Modelling* **16:** 121–136.

———. 1993. Evolutionary history of reverse transcriptase. In *Reverse transcriptase* (ed. A.M. Skalka and S.P. Goff), pp. 425–444. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

McClure, M.A., M.S. Johnson, D.F. Feng, and R.F. Doolittle. 1988. Sequence comparisons of retroviral proteins: Relative rates of change and general phylogeny. *Proc. Natl. Acad. Sci.* **85:** 2469–2473.

McDonald, J.F. 1993. Evolution and consequences of transposable elements. *Curr. Opin. Genet. Dev.* **3:** 855–864.

———. 1998. Transposable elements, gene silencing and macroevolution. *Trends Ecol. Evol.* **13:** 94–95.

Miki, Y. 1998. Retrotransposal integration of mobile genetic elements in human diseases. *J. Hum. Genet.* **43:** 77–84.

Miller, W.J., L. Kruckenhauser, and W. Pinsker. 1996. The impact of transposable elements on genome evolution in animals and plants. In *Transgenic Organisms-Biological and Social Implications* (ed. J. Tomiuk, K. Woerhm, and A. Sentker), pp. 21–34. Birkhauser Verlag, Basel, Switzerland.

Molloy, S.S., P.A. Bresnahan, S.H. Leppla, K.R. Klimpel, and G. Thomas. 1992. Human furin is a calcium-dependent serine endoprotease that recognizes the sequence Arg-X-X-Arg and efficiently cleaves anthrax toxin protective antigen. *J. Biol. Chem.* **267:** 16396–16402.

Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heijne. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10:** 1–6.

Page, R.D. 1996. TreeView: An application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12:** 357–358.

Rost, B., R. Casadio, P. Fariselli, and C. Sander. 1995. Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4:** 521–533.

SanMiguel, P., A. Tikhonov, Y.K. Jin, N. Motchoulskaia, D. Zakharov, A. Melake-Berhan, P.S. Springer, K.J. Edwards, M. Lee, Z. Avramova, and J.L. Bennetzen. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274:** 765–768.

SanMiguel, P., B.S. Gaut, A. Tikhonov, Y. Nakajima, and J.L. Bennetzen. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20:** 43–45.

Solovyev, V.V., A.A. Salamov, and C.B. Lawrence. 1994. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* **22:** 5156–5163.

Swofford, D.L. 1999. *PAUP*: Phylogenetic analysis using parsimony (* and other methods)*. Sinauer Associates, Sunderland, MA.

Temin, H.M. 1985. Reverse transcription in the eukaryotic genome: Rretroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Mol. Biol. Evol.* **2:** 455–468.

Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876–4882.

Xiong, Y. and T.H. Eickbush. 1988. Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. *Mol. Biol. Evol.* **5:** 675–690.

———. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9:** 3353–3362.

Xiong, Y., W.D. Burke, and T.H. Eickbush. 1993. Pao, a highly divergent retrotransposable element from Bombyx mori containing long terminal repeats with tandem copies of the putative R region. *Nucleic Acids Res.* **21:** 2117–2123.

Yoder, J.A., C.P. Walsh, and T.H. Bestor. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13:** 335–340.

Received May 12, 1999; accepted in revised form August 12, 1999.